DOCUMENT RESUME

ED 408 303 TM 026 505

AUTHOR Thompson, Bruce; Snyder, Patricia A.
TITLE Use of Statistical Significance Tests and Reliability Analyses in Published Counseling Research.
PUB DATE 25 Mar 97
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Counseling; Educational Research; *Effect Size; Evaluation Methods; Reliability; Research Methodology; *Research Reports; *Scholarly Journals; Scores; *Statistical Significance; *Test Use
IDENTIFIERS *Journal of Counseling and Development; Research Replication

ABSTRACT

The mission of the "Journal of Counseling and Development" (JCD) includes the attempt to serve as a "scholarly record of the counseling profession" and as part of the "conscience of the profession." This responsibility requires the willingness to engage in self-study. This study investigated two aspects of research practice in 25 quantitative studies reported in 1996 JCD issues, the use and interpretation of statistical significance tests, and the meaning of and ways of evaluating the score reliabilities of measures used in substantive research inquiry. Too many researchers have persisted in equating result improbability with result value, and too many have persisted in believing that statistical significance evaluates result replicability. In addition, too many researchers have persisted in believing that result improbability equals the magnitude of study effects. Authors must consistently begin to report and interpret effect sizes to aid the interpretations they make and those made by their readers. With respect to score reliability evaluation, more authors need to recognize that reliability inures to specific sets of scores and not to the test itself. Thirteen of the JCD articles involved reports of score reliability in previous studies and eight reported reliability coefficients for both previous scores and those in hand. These findings suggest some potential for improved practice in the quantitative research reported in JCD and improved editorial policies to support these changes. (Contains 39 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 408 303

jcd.wp1 3/7/97

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
BRUCE THOMPSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

USE OF STATISTICAL SIGNIFICANCE TESTS AND RELIABILITY ANALYSES IN PUBLISHED COUNSELING RESEARCH

Bruce Thompson
Texas A&M University
and
Baylor College of Medicine

Patricia A. Snyder
LSU Medical Center

BEST COPY AVAILABLE

TM026505

Paper presented at the annual meeting of the American Educational Research Association (*Session #13.07*), Chicago, March 25, 1997.

ABSTRACT

The mission of the Journal of Counseling and Development (JCD) includes serving as "a scholarly record of the counseling profession" and as part of "conscience of the profession." This ambitious responsibility may require the willingness to engage in occasional self-study. The present study investigated two aspects of research practice in the quantitative studies reported in 1996 JCD issues.

The Journal of Counseling and Development (JCD) has been described as "a scholarly record of the counseling profession" (Borders, 1996, p. 3). JCD is distributed to all members of the American Counseling Association and to additional individual and institutional subscribers. Because nearly 60,000 copies of each issue are published, JCD is well positioned to move the field, and ultimately to impact the care that clients receive.

But the journal's impact is influenced by more than circulation figures--the journal's mission itself also affects impact. As described by Borders (1996), JCD's mission is to address "the needs, concerns, and interests of members of the American Counseling Association... As a scholarly resource for a diverse readership, the Journal, along with other voices of ACA, also serves in part as the conscience of the profession" (p. 3).

Such an ambitious mission recognizes that JCD articles can affect clinical practice, and thus the editorial policies and practices of the journal must be exercised responsibly. Part of this responsibility may involve the willingness to engage in occasional self-study.

The present paper represents an effort by one JCD editorial board member to study two aspects of research practice in the quantitative studies reported in recent JCD issues. As Sexton (1996) noted in a recent JCD article, "For [counseling] research to be useful, studies must incorporate the basic elements of an appropriate research design, relevant measures, and methodological improvements currently advocated in the literature" (pp. 598-599).

Because both the counseling and methodological fields do over time evolve changing consensus about what constitutes accepted practice, it is important to self-evaluate contemporary practice on a regular basis, to insure that on-going practice reflects current thinking.

The present work first describes contemporary thinking regarding two methodological issues: (a) the use and interpretation of statistical significance tests, and (b) the meaning of and ways to evaluate the score reliabilities of the measures used in substantive quantitative inquiry. Next, practice as regards these two issues within recent JCD articles is described. Finally, some recommendations for potential improvement are presented.

Two Evolving Elements of Methodological Practice

There have certainly been several areas of methodological practice in which thinking about acceptable practice has evolved. But statistical significance testing and the evaluation of score reliability have been among the arenas in which especially noteworthy changes have been occurring.

Statistical Significance Testing

Social scientists have expressed increasing concerns regarding the use and interpretation of statistical significance tests. Essentially, the social sciences have been moving away from emphasizing statistical significance tests and toward emphasizing evaluations of (a) practical significance and (b) result replicability. The concerns underlying these views can only be briefly summarized here (but see Kirk (1996), Schmidt (1996) and Thompson (1996)). Three recent developments in discussions

surrounding the use and interpretation of statistical significance tests reflect these underlying concerns:

1. After prolonged deliberation over the course of the last two years, the APA Board of Scientific Affairs recently named a Task Force on Statistical Inference (Azar, 1997; Shea, 1996). The Task Force is charged with recommending policies and practices leading to more informed and thoughtful statistical analyses, particularly as regards statistical significance testing.
2. The new fourth edition of the American Psychological Association (APA) style manual (APA, 1994) included an important, but largely unheralded, shift in APA editorial policy regarding the use of statistical significance testing in quantitative research. The manual noted that:

   > Neither of the two types of probability values reflects the importance or magnitude of an effect *because both depend on sample size*... You are encouraged to provide effect-size information. (APA, 1994, p. 18, emphasis added)

3. A series of recent articles have been published as regards these concerns. For example, the <u>American Psychologist</u> published a seemingly periodic series of articles on the limits of statistical significance testing (cf. Cohen, 1990, 1994; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989). An entire 1993 issue (vol. 61, no. 4) of the <u>Journal of Experimental Education</u> was devoted to these themes. [Less

recent but influential works in this genre have been the publications by Rozeboom (1960), Morrison and Henkel (1970), Carver (1978), Meehl (1978), Shaver (1985).] The articles by Carver (1978) and Cohen (1994) have been particularly influential.

These concerns and others have arisen because some researchers have not understood what their p calculated values actually evaluate (Carver, 1978). Thompson (1996, p. 27) summarized what p really tests: "$p_{CALCULATED}$ is the probability (0 to 1.0) of the sample statistics, given the sample size, and assuming the sample was derived from a population in which the null hypothesis (H_0) is exactly true." Thompson (1996) and Shaver (1993) provided further explanation regarding what statistical significance tests evaluate.

But three myths have persisted, notwithstanding the availability of such explanations. These myths have been particularly influential, partly because the myths have been adopted unconsciously, so that most researchers cannot readily scrutinize the premises underlying their behavior.

First, too many researchers have persisted in *equating result improbability with result value*. But an unlikely event simply is not necessarily an important event. Shaver's (1985, p. 58) classic hypothetical dialogue between two teachers illustrates the folly of equating improbability with importance:

> Chris: ...I set the level of significance at .05, as my advisor suggested. So a difference that large would occur by chance less than five times in a

hundred if the groups weren't really different. An unlikely occurrence like that *surely* must be important.

Jean: Wait a minute, Chris. Remember the other day when you went into the office to call home? Just as you completed dialing the number, your little boy picked up the phone to call someone. So you were connected and talking to one another without the phone ever ringing... Well, that must have been a truly important occurrence then?

As Thompson (1993b, p. 365) explained, "If the computer package did not ask you your values prior to its analysis, it could not [possibly] have considered your value system in calculating p's, and so p's cannot be blithely used to infer the value of research results."

Second, too many researchers have persisted in believing that *statistical significance evaluates result replicability*. Testing the probability that sample results were descriptive of the population would bear upon replicability, since future samples from that population should involve comparable results. But statistical significance tests do not test the population; instead, they do the opposite--they *assume* specified population parameters, and test the sample probability (Cohen, 1994; Thompson, 1996)!

Third, too many researchers have persisted in believing that *result improbability equals the magnitude of study effects*. It is true that effect sizes (e.g., eta^2, $omega^2$, r^2, Cohen's d) can be

computed in all studies (see Kirk (1996) or Snyder and Lawson (1993)). And it is also true that effect sizes do affect $p_{CALCULATED}$ values and that, all things equal, larger effects will result in smaller $p_{CALCULATED}$ values.

However, a study's $p_{CALCULATED}$ values are influenced by at least seven interrelated study features (Schneider & Darcy, 1984), and not only by effect size magnitudes. For example, the reliability of the scores in hand itself impacts $p_{CALCULATED}$. Thus, a $p_{CALCULATED}$ value cannot be employed as a pure measure of effect.

Sample size is a very big influence on whether or not results are statistically significant. This means that "virtually any study can be made to show [statistically] significant results if one uses enough subjects" (Hays, 1981, p. 293). Similarly, Nunnally (1960, p. 643) noted that, "If the null hypothesis is not rejected, it is usually because the N is too small."

All this means that statistical significance testing can become largely a test of what we already know: our sample size. As Thompson (1992) observed,

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they're tired. This tautology has created considerable damage as regards the cumulation of

knowledge... (p. 436)

Score Reliability Evaluation

Readers commonly encounter research in which authors talk about "the reliability of the test," or in which a given test is described as being reliable. Counseling researchers need to understand that such telegraphic ways of speaking inherently assert untruths. Put simply, *reliability is a characteristic of scores for the data in hand*, and not of a test per se. Unfortunately, these habits of speaking and writing are not merely sloppy:

> This is not just an issue of sloppy speaking--the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice. (Thompson, 1992, p. 436)

As Rowley (1976, p. 53, emphasis added) noted, "It needs to be established that an instrument itself is neither reliable nor unreliable.... A single instrument can produce *scores* which are reliable, and other *scores* which are unreliable." Similarly, Crocker and Algina (1986, p. 144, emphasis added) explained that, "...A test is not 'reliable' or 'unreliable.' Rather, reliability is a property of the *scores* on a test for a *particular* group of examinees." In another widely respected text, Gronlund and Linn (1990, p. 78, emphasis in original) noted,

> Reliability refers to the *results* obtained with an evaluation instrument and not to the instrument

> itself.... Thus, it is more appropriate to speak of the reliability of the "test scores" or of the "measurement" than of the "test" or the "instrument."

The participants in research themselves affect the reliability of scores, and thus it is incongruous to speak of "the reliability of the test" without considering to whom the test was administered, and other facets of the measurement protocol. Reliability estimates are driven by variance--typically, greater score variance leads to greater score reliability, and so more *heterogeneous* samples often lead to more *variable* scores, and thus to higher reliability. Therefore, the same measure, when administered to more heterogenous or to more homogeneous sets of participants, will yield scores that result in different reliability estimates.

One implication of these realizations is that *we ought to confirm the reliability of our own scores in each of our substantive studies*. As Dawis (1987, p. 486) observed, "...Because reliability is a function of sample as well as of instrument, it should be evaluated on a sample from the intended target population--an obvious but sometimes overlooked point."

However, a disturbing proportion of researchers fail even only to report (a) the reliability of scores in previous studies on the measures being used in substantive inquiries, and (b) explicit comparisons of the samples and test conditions in their studies with the samples and test conditions involved in previous reliability studies.

For example, with respect to the American Educational Research Journal, Willson (1980) reported that:

> ...Only 37% of the AERJ studies explicitly reported reliability coefficients for the data analyzed. Another 18% reported only indirectly through reference to earlier research.... That reliability... is unreported in almost half the published research is... inexcusable at this late date...." (pp. 8-9)

A more recent "perusal of contemporary psychology journals demonstrates that quantitative reports of scale reliability and validity estimates are often missing or incomplete" (Meier & Davis, 1990, p. 113); and that "the majority [95%, 85% and 60%] of the scales described in the [three Journal of Counseling Psychology] JCP volumes [1967, 1977 and 1987] were not accompanied by reports of psychometric properties" (p. 115).

The concern for score reliability in substantive inquiry is not just some vague statistician's nit-picking. Score reliability directly (a) affects our ability to achieve statistically significance and (b) attenuates the effect sizes for the studies we conduct. For example, even if the "true" relationship between perfectly reliable measures of X and Y was perfect (i.e., $\hat{r}_{XY}^2$ = 1.0), the detectable effect in any study can never exceed the product of the reliability coefficients for the two sets of scores: max r_{XY} = the square root of (r_{XX} * r_{YY}) (Locke, Spirduso & Silverman, 1993, p. 17). It certainly may be important to consider

such effects as part of result interpretation, once the study has been conducted.

Practice Within *JCD*

For the purposes of the present study, all the quantitative studies in the 1996 issues of the Journal of Counseling and Development were analyzed. Twenty-six quantitative studies reporting statistical tests were identified [in citing the studies here, only volumes and page numbers are cited, rather than inserting all the studies into the references, since interested readers can still readily locate the studies]. One of the 26 studies (Elliott, Scewchuk, Richeson, Pickelman & Franklin, vol. 74, pp. 645-651) was excluded, since an important but hybrid application of statistical testing was employed (i.e., structural equation modeling) in which the researcher seeks to not reject the null hypothesis, and often modifies the hypothesis until this result is achieved.

For each study for all statistical significance tests a variance-accounted-for effect size analogous to r^2 was computed, using the procedures described by Kirk (1996) and Snyder and Lawson (1993). This resulted in the computation of an average of 10.9 effect sizes per study (SD = 13.4) for a total of 274 effects. However, it must be noted that some authors only reported results for statistically significant effects, and effect sizes could not be computed when authors provided no information except that certain results were not statistically significant.

Additionally, interpretations of statistical significance were

noted and it was recorded which, if any, studies reported effect sizes. The studies were also each characterized as regards the analysis of score reliability. Studies were assigned to three categories: (a) no reliability coefficients were reported for the measures used; (b) only reliability coefficients from previous studies were reported for the measures; or (c) reliability coefficients were reported for the scores actually being subjected to substantive analysis in the study.

Results

Statistical Significance Testing

In 15 of the 25 studies authors reported at least one effect size (e.g., r^2, eta^2, $omega^2$, Cohen's $\underline{d}$). But here these reports invariably took the form of squared correlation coefficients (e.g., r^2, multiple R^2, or canonical R_C^2). Some such reports were seemingly incidental to the primary purpose of the studies (e.g., values in a bivariate correlation matrix were presented, but the primary focus of the study involved tests of mean differences).

In only 2 of the 25 studies did authors report effect sizes and interpret them as such (i.e., evaluated their relative magnitudes). In the first of these two studies, Perosa (vol. 74, no. 4) reported a statistically significant squared canonical correlation (R_C^2 = 23%), but noted that this "explains only a small percentage of the variance" (p. 390) and characterized this as "a low amount of variance accounted for" (p. 390).

Of course, interpreting variance-accounted-for is a professional judgment, and different people may reasonably reach

different judgments regarding a given result. However, many researchers (cf. Cohen, 1988) might regard 23% as a large effect size. In fact, for the 274 effects reported in the 25 JCD articles, the mean of calculated variance-accounted-for effects was .148 (SD = .134).

The mean of the effects within each of the 25 studies was also computed; the mean of these 25 within-study effect sizes was .168 (SD = .083). These 25 within-study mean effects ranged from .037 to .399. However, it should be noted that the study with a within-study mean effect size of .399 was a validity study (Melchert, Hays, Wiljanen & Kolocek, vol. 74, pp. 640-644) in which large effect sizes should be expected.

In a second study, Rice and Cummins (vol. 75, pp. 50-57) reported a series of effect sizes in the form of R^2 values. These results were not only reported, but were explicitly interpreted as variance-accounted-for statistics. Furthermore, the authors explicitly compared their effects with those reported in previous studies, including previous meta-analytic work.

In one additional study (Kaminski & McNamara, vol. 74, pp. 288-294) the authors did not report or interpret effect sizes as such, but clearly did distinguish statistical from practical significance. In their discussion section these authors computed and interpreted statistics completely separate from those in their results section; these computations involved counts of clients showing clinically significant improvements.

Several authors referred to "significant" results when the

intent was apparently to refer to "statistically significant" results: "there were significant differences between Black and White women" (Carter & Parks, vol. 74, p. 486); "none of the masculinity-related variables was significantly correlated with men's self-reported likelihood of raping, nor did variables combine to predict a significant amount of variance" (Truman, Tokar & Fischer, vol. 74, p. 560); "women... tended to perceive significantly more barriers" (Luzzo & Hutcheson, vol. 75, p. 128); and "women's experience of inequities... may be significantly associated with negative self-estimates" (Ancis & Phillips, vol. 75, p. 135). Such language is fairly common, and is not proscribed by style manuals.

However, some have recommended that the phrase "statistically significant" should always be used when referring to statistical tests (cf. Thompson, 1996). For example, Carver (1993) noted:

> When trying to emulate the best principles of science, it seems important to say what we mean and to mean what we say. Even though many readers of scientific journals know that the word *significant* is supposed to mean *statistically significant* when it is used in this context, many readers do *not* know this. Why be unnecessarily confusing when clarity should be most important? (p. 288, emphasis in original)

In any case, it can be confusing if "significant" is used within the same article in some places to mean "important" and in other

places to mean "statistically significant" (see Kaminski & McNamara, vol. 74, pp. 288-294).

Score Reliability Evaluation

One of the 25 studies involved the creation of a new measure, and an alpha coefficient was reported for the data actually being analyzed in the published report. In an additional eight articles authors reported reliability coefficients both for their own data and for scores in previous studies.

In another 13 articles, authors reported specific reliability coefficients for previous studies' scores on the measures they used. However, these authors did not make explicit comparisons of the participants or measurement features in the previous studies with their own. This pattern may stem from an unconscious belief that "tests are reliable," and that therefore reliability is always assured whenever certain measures are employed.

Some authors explicitly invoked language asserting that tests are reliable. Examples included: "the BES is highly reliable" (Kaminski & McNamara, vol. 74, p. 289); "Cronbach's alpha for the DAS is .96" (Contreras, Hendrick & Hendrick, vol. 74, p. 410); "weak reliabilities of the Preencounter and Encounter subscales" (Carter & Parks, vol. 74, p. 488); "reliabilities of TRIG subscales" (Brown, Richards & Wilson, vol. 74, p. 506); and "may be a reliable and valid measure" (Melchert, Hays, Wiljanen & Kolocek, vol. 74, p. 642).

Discussion

The present inquiry focused on two aspects of analytic

practice in the quantitative articles recently published in the Journal of Counseling and Development. As regards *statistical significance testing*, in two of the 25 articles authors reported and interpreted effect sizes. In an additional study, effect sizes were not reported in the results section, but the distinction between statistical and practical significance was clearly drawn in the discussion. The articles routinely described results as "significant," and not as "statistically significant," as some have recommended (e.g., Carver, 1993).

Of course, it is possible that some of the 1996 articles were written only shortly after the 1994 APA style manual, which at least "encourages" the reporting of effect sizes, was first published. And certainly the manuscripts were written before the creation of the new APA Task Force on Statistical Inference, and before the publication of some of the more recent works on this topic (e.g., Kirk, 1996; Schmidt, 1996; Thompson, 1996).

Authors must consistently begin to report and interpret effect sizes, to aid the interpretations made both by themselves and independently by their readers. Reporting effect sizes also assists in the meta-analytic synthesis of diverse research findings (e.g., findings regarding validation of counseling interventions). One 1996 JCD author made this point quite directly:

> The interpretation of results is more complex than noting the statistically significant findings. Traditional statistical interpretation has been criticized because statistical significance does not

> guarantee that the results have clinical or practical meaning. (Whiston, 1996, p. 619)

As regards *score reliability evaluation*, more authors need to recognize that reliability inures to specific sets of scores, and not to the test itself. It would help more readers recognize this truism if authors regularly referred to the reliability of scores, and refrained from language implying that tests are reliable.

An impressive number of articles involved reports of score reliability in previous studies (n=13), while one article involved reporting of reliability results for the scores in hand, and eight of the 25 articles reported reliability coefficients for both previous studies and for the scores in hand actually being subjected to substantive analysis. This result compares extremely favorably with practices in some psychology journals (e.g., Snyder & Thompson, 1997), and even with practice in measurement journals (Thompson, 1994b).

These findings and the discussion herein suggest some potential for improved practice in reporting and interpretation within the quantitative research published in JCD. First, it might be appropriate to ask all authors to report effect sizes in conjunction with their statistical significance tests. The American Counseling Association is not bound to adhere exactly to admonitions in the 1994 APA style manual, which only "encourage" such reports. It may be reasonable for the Journal of Counseling and Development to articulate additional expectations, just as other journals have done (e.g., Heldref Foundation, in press;

Thompson, 1994a).

Second, more authors might be encouraged to report score reliability for their own data, even though many JCD authors do so already. Such reports may affect the interpretation of the effects reported in substantive studies, since measurement error tends to attenuate effect sizes. Futhermore, such reports also lead to the more rapid cumulation of evidence about the psychometric properties of scores from various measures across variations in samples and measurement contexts. Authors might also be encouraged to report explicit comparisons of the participants or measurement features in the previous studies with their own, when reliability coefficients from previous studies are reported.

The Journal of Counseling and Development is read by nearly 60,000 counselors, who presumably consult the journal to improve their work with clients, and who view the journal as part of "the conscience of the profession" (Borders, 1996, p. 3). The practices recommended here may further facilitate the journal's pursuit of its important mission.

References

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Azar, B. (1997). Apa task force urges a harder look at data. The APA Monitor, 28(3), 26.

Borders, L.D. (1996). The *Journal of Counseling & Development*: On its purpose, function, and goals. Journal of Counseling and Development, 75, 3-4.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cohen, J. (1988). Statistical power analysis (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Dawis, R.V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-489.

Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

Heldref Foundation. (in press). Guidelines for contributors. Journal of Experimental Education.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56(5), 746-759.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Locke, L.F., Spirduso, W.W., & Silverman, S.J. (1993). Proposals that work: A guide for planning dissertations and grant proposals (3rd ed.). Newbury Park, CA: SAGE.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Meier, S.T., & Davis, S.R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. Journal of Counseling Psychology, 37, 113-115.

Morrison, D.E., & Henkel, R.E. (Eds.). (1970). The significance test controversy. Chicago: Aldine.

Nunnally, J. (1960). The place of statistics in psychology. Educational and Psychological Measurement, 20, 641-650.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. American Psychologist, 46, 1086-1087.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and

the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Rowley, G.L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1(2), 115-129.

Schneider, A.L., & Darcy, R.E. (1984). Policy implications of using significance test in evaluation research. Evaluation Review, 8, 573-582.

Sexton, T.L. (1996). The relevance of counseling outcome research: Current trends and practical implications. Journal of Counseling and Development, 74, 590-600.

Shaver, J. (1985). Chance and nonsense. Phi Delta Kappan, 67(1), 57-60.

Shaver, J. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61(4), 293-316.

Shea, C. (1996). Psychologists debate accuracy of "significance test." Chronicle of Higher Education, 42(49), A12, A16.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61(4), 334-349.

Snyder, P.A., & Thompson, B. (1997, January). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested

alternatives. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Document Reproduction Service No. ED forthcoming)

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1993b). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61(4), 361-377.

Thompson, B. (1994a). Guidelines for authors. Educational and Psychological Measurement, 54(4), 837-847.

Thompson, B. (1994b, January). It is incorrect to say "The test is reliable": Bad language habits can contribute to incorrect or meaningless research conclusions. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX. (ERIC Document Reproduction Service No. ED 367 707)

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Whiston, S.C. (1996). Accountability through action research: Research methods for practitioners. Journal of Counseling and Development, 74, 616-623.

Willson, V.L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9(6), 5-10.